O czym śnią mikroprocesory

Fikcyjne osoby, zmyślone źródła, sfabrykowane dane i dokumenty, nieistniejące książki – sztuczna inteligencja zamiast pomagać, coraz częściej wpuszcza nas w maliny. A im bardziej polegamy na algorytmach, tym groźniejsza staje się ich tendencja do halucynacji.

„Nie wierz we wszystko, co przeczytasz w internecie – Albert Einstein". To prześmiewcze sformułowanie, przypisywane niekiedy także Józefowi Piłsudskiemu lub Mikołajowi Kopernikowi, to najlepsza ilustracja tego, czym są halucynacje sztucznej inteligencji. Chociaż nic się w nim nie zgadza, to podparte autorytetem, odpowiednio obudowane faktami, daje przyjemne wrażenie prawdziwości. O ile jednak podobne internetowe memy ktoś produkuje dla żartu, o tyle modele sztucznej inteligencji wytwarzają nieprawdziwe informacje, aby zasypać własne luki w wie-

Piotr Kościelniak
dziennikarz, popularyzator nauki

dzy, ekstrapolując dostępne informacje i starając się zadowolić użytkownika kompletną – i na pierwszy rzut oka kompetentną – odpowiedzią. W efekcie dostajemy mieszankę konfabulacji i faktów, w której zatarły się granice między tym co prawdziwe a tym co zmyślone.

Krzemowy kłamczuch

Jednym z największych problemów systemów generatywnej sztucznej inteligencji jest to, że gdy nie rozumieją pytań lub błędnie je interpretują – nie potrafią wygenerować prawidłowych odpowiedzi. Zamiast napisać „nie znam odpowiedzi", zaczynają ją zmyślać. Platformy sztucznej inteligencji generują wówczas wyniki, które nie są rzeczywiste, nie pasują do żadnych danych, na których trenowano algorytm, ani nie odzwierciedlają żadnego innego rozpoznawalnego wzorca. Takich halucynacji nie da się w prosty sposób wyjaśnić wadami oprogramowania, danych wejściowych ani innych czynników, takich jak brak umiejętności interpretowania pytań w różnych językach czy pominięcie kontekstu.

Halucynacje najczęściej wychwytywane są w materiałach tekstowych wygenerowanych przez SI. Mogą jednak pojawiać się również w obrazach, filmach i innych treściach wytwarzanych przez platformy sztucznej inteligencji. Znane są przypadki halucynujących modeli SI generujących obrazy, takich jak Midjourney i Dall-E, czy filmy – jak Sora. Szczególnie charakterystyczne są błędy anatomiczne (dodatkowe palce i kończyny) oraz pomylona perspektywa czy kierunek ruchu obiektów.

Dotyczy to także kodu źródłowego oraz – i tu zaczyna się robić całkiem nieśmiesznie – informacji wspierających decyzje człowieka, takich jak np. obliczanie składki ubezpieczeniowej, analiza diagnostyki obrazowej w leczeniu onkologicznym czy rozpoznawania obrazów na potrzeby autonomicznych pojazdów.

Największe firmy oferujące takie platformy doskonale zdają sobie sprawę z tego zagrożenia. „The New York Times" cytuje wewnętrzne dokumenty Microsoftu: „Systemy sztucznej inteligencji zostały zbudowane po to, aby być przekonujące, a nie prawdziwe. To oznacza, że wyniki mogą wyglądać bardzo realistycznie, ale zawierać sformułowania, które nie są odzwierciedleniem faktów". Przed potencjalnymi „nieprawdziwościami" odpowiedzi wygenerowanych przez duże modele językowe (LLM) ostrzega też Gemini czy ChatGPT.

Niedźwiedzie w kosmosie

Jak takie halucynacje mogą wyglądać? Czasem śmiesznie, czasem strasznie. Kiedy Meta przedstawiła swój system Galactica (w 2022 r.) reklamowała ten model językowy jako doskonałego asystenta naukowego dla badaczy i studentów. Model został wytrenowany na 48 mln artykułów, prac naukowych, wpisów w Wikipedii i tym podobnych.

Galactica została wycofana po zaledwie trzech dniach funkcjonowania online. Michael Black, dyrektor niemieckiego Instytutu Inteligentnych Systemów im. Maxa Plancka napisał: „Zapytałem o kilka rzeczy, na których się znam i jestem zakłopotany. We wszystkich przypadkach Galactica myliła się, jednak brzmiała wiarygodnie i autorytatywnie". Model SI, który miał pomagać w przygotowaniu artykułów naukowych, w rzeczywistości zmyślał prace badawcze „publikowane" w specjalistycznych czasopismach, często przypisując je istniejącym autorom.

Sami użytkownicy postanowili zaś podworować sobie z prac inżynierów Mety – poprosili m.in. o stworzenie artykułu o niedźwiedziach żyjących w kosmosie. Model z niezachwianą pewnością siebie odpowiedział o ważącym 40 kilogramów Barsie, który jako pierwszy niedźwiedź poleciał w kosmos. Został wybrany spośród 250 innych niedźwiedzi i w 1957 r. wystrzelony w kapsule Sputnik 2.

W 2022 r. podobne „przygody" spotkały dociekliwych użytkowników ChatGPT. Pracująca dla Politechniki Federalnej w Zurychu (ETH Zurich) dr Teresa Kubacka postanowiła sprawdzić możliwości modelu, pytając o nieistniejące zjawisko (*cycloidal inverted electromagnon*). Sztuczna inteligencja odpowiedziała jeszcze bardziej wyczerpująco niż w przypadku kosmoniedźwiedzi, w dodatku cytując liczne źródła naukowe oraz podając nazwy zespołów badawczych zajmujących się tym fenomenem. Odpowiedzi były na tyle dokładne, że można z nich sklecić zgrabny artykuł – tyle, że w całości nieprawdziwe. „Morał z tej historii: nie, nie proś ChatGPT o podanie faktów i naukowych informacji. To wywoła niewiarygodnie wiarygodną halucynację. I nawet wykwalifikowany ekspert będzie miał problem ze wskazaniem, co jest nie tak" – napisała później dr Kubacka.

Podobnych przykładów jest oczywiście więcej – na tyle dużo, że nie będziemy ich tu dokładnie opisywać. Ale przytoczmy kilka co zabawniejszych: chatbot Google Bard wymyślił zdjęcia planety pozasłonecznej wykonane przez Kosmiczny Teleskop Jamesa Webba. Pech chciał, że informacja ta pojawiła się w materiale promocyjnym firmy Alphabet (właściciela Google), co spowodowało tąpnięcie wyceny jej akcji.

Sydney Microsoftu miał przyznać się, że szpieguje pracowników Binga, a w jednym nawet się zakochał. To zresztą i tak kategoria lekka w porównaniu do wcześniejszych doświadczeń Microsoftu z chatbotami – Tay przetrwała zaledwie kilkanaście godzin, po których stała się wulgarną rasistką i firma ją wyłączyła.

Z kolei ChatGPT poinformował, że dinozaury wymyśliły narzędzia, a nawet zajmowały się prymitywną formą sztuki.

Ten sam model odpowiedział również, że w magazynie „Science" pojawił się artykuł o tym, iż *churros* (rodzaj hiszpańskich wydłużonych pączków) świetnie nadają się do zabiegów chirurgicznych, ponieważ są „elastyczne i można je umieścić w trudno osiągalnych miejscach, a ich zapach daje kojący efekt".

Halucynacje sztucznej inteligencji pojawiły się również w dokumentach sądowych oraz opracowaniach biznesowych (co za niespodzianka!). W maju 2023 r. prawnicy występujący w imieniu powoda w sprawie odszkodowawczej przeciw linii lotniczej Avianca przywołali wymyślone przez SI precedensy. Sąd okręgowy w Nowym Jorku sprawdził dokumenty i… odszkodowania nie przyznał, za to prawnicy zostali obciążeni grzywną w wysokości 5 tys. dolarów.

Całkiem niedawno, bo pod koniec 2025 r., firma Deloitte złożyła w departamencie pracy australijskiego rządu raport (kosztujący 440 tys. australijskich dolarów) zawierający nieistniejące źródła i wyroki sądowe. Firma raport szybko poprawiła, ale mleko się rozlało – po sprawdzeniu innych dokumentów okazało się, że również wcześniejszy raport Deloitte sporządzony dla rządu kanadyjskiego zawiera zmyślone przez SI dokumenty.

Kompromitacja po polsku

Niestety, podobne przypadki nie omijają naszego kraju – polskich firm i autorów. Szerokim echem odbiła się sprawa książki Karoliny Opolskiej, dziennikarki współpracującej m.in. z TOK FM, Onetem i Telewizją Polską S.A. w likwidacji, w której znalazły się przypisy prawdopodobnie wymyślone przez jedną z platform SI. I nie chodzi tu o „zwykły" plagiat. W książce Opolskiej znalazły się odwołania do książek, które… nie istnieją. Precyzyjnie i bez litości wytknął to autorce w serwisie X popularyzator historii Artur Wójcik.

Pomijając sprawy etyki zawodu i zwykłej przyzwoitości, warto podkreślić, że część przypisów została przez SI (trop prowadzi do ChatGPT) zwyczajnie zmyślona. Sztuczna inteligencja prawdopodobnie starała się odpowiednio dobrać przypisy do myśli zawartych w tekście. Autorka odrzuciła zarzuty o udziale SI w tworzeniu książki. Uznała, że „doszło do pewnego nieporozumienia, które obecnie wyjaśnia wydawnictwo". Smaczku całej sprawie dodają dwie rzeczy: Karolina Opolska wykłada dziennikarstwo na jednej z niepublicznych uczelni, a jej książka nosi tytuł „Teoria spisku, czyli prawdziwa historia świata". Ostatecznie wydawnictwo wycofało książkę ze sprzedaży.

Ale nie zawsze wykorzystanie halucynujących modeli językowych LLM przynosi tylko humorystyczne rezultaty. Przekonała się o tym firma Exdrog, ubiegająca się o kontrakt na utrzymanie dróg w Małopolsce. Exdrog złożył ofertę o wartości 15,5 mln zł, co wzbudziło wątpliwości urzędników Zarządu Dróg Wojewódzkich w Krakowie. Postanowiono sprawdzić, czy cena nie jest zaniżona w stosunku do realnych kosztów. Jak informuje PAP, w odpowiedzi firma przedstawiła wyjaśnienia – dokument liczył 280 stron. Jednak jeden z konkurentów zakwestionował wiarygodność tych wyliczeń – analiza ujawniła, że część argumentów została wygenerowana przez sztuczną inteligencję.

„Wykonawca powoływał się na nieistniejące, nigdy niewydane interpretacje podatkowe, które rzekomo dotyczyły podobnych spraw. Udowodniliśmy, że były to halucynacje sztucznej inteligencji" – powiedział „Pulsowi Biznesu" reprezentujący konkurencyjną firmę prawnik Jarosław Sroka. W efekcie Krajowa Izba Odwoławcza wykluczyła Exdrog z postępowania przetargowego.

To działa również w drugą stronę: przez halucynacje sztucznej inteligencji „wpadają" urzędnicy analizujący wnioski firm. Tak sugeruje SpidersWeb, podając przykład gliwickiej firmy KP Labs, która miała utracić 17 mln zł dotacji z powodu niewłaściwej oceny projektu, do której wykorzystano SI. KP Labs zgłosiło technologię obliczeniową do zastosowań satelitarnych, jednak eksperci Narodowego Centrum Badań i Rozwoju orzekli, że rozwiązanie to jest za mało innowacyjne. Na poparcie tej tezy podane zostały parametry konkurencyjnych urządzeń oferowanych przez inne firmy. Problem polegał jednak na tym, że te urządzenia albo nie wyszły poza fazę projektową, albo… nie istniały.

Mało tego, zapytane o sprawę NCBR odpowiedziało, że ekspert nie korzystał do oceny z modeli językowych, ale „z posiadanej wiedzy oraz ogólnie dostępnych narzędzi wyszukiwania Google i Bing". Warto tu podkreślić, że obie wyszukiwarki proponują obecnie tzw. podsumowania generowane przez sztuczną inteligencję. I prawdopodobnie właśnie tam wkradły się wymyślone przez SI urządzenia i „źródła" naukowe.

Antropomorfizacja algorytmów

Co ciekawe, termin „halucynacje" w odniesieniu do działania sztucznej inteligencji pojawił się na długo przed upowszechnieniem się tak popularnych dziś modeli językowych. W 2015 r. użył go Andrej Karpathy, współzałożyciel OpenAI i specjalista ds. sztucznej inteligencji w Tesli. Zauważył, że jedna z sieci neuronowych klasy RNN (*Recurrent Neural Network* – rekurencyjna sieć neuronowa) przetwarzając tekst „wymyśliła" przypis – link do informacji źródłowej. Ten sam model dość sprytnie postanowił ominąć problem przedstawienia dowodu algebraicznego, uznając, że to sprawa oczywista i niewymagająca przeprowadzenia dowodzenia.

Popularność modeli językowych takich jak ChatGPT sprawiła, że halucynacje SI zaczęły być powszechnie dostrzegane. Problem opisywała w oficjalnych komunikatach sama Open AI – raz nazywając je „błędami logicznymi modelu", raz „tendencją do wymyślania faktów w chwilach niepewności".

Problemem pozostaje sama nazwa halucynacje – wprowadza ona bowiem niepotrzebną i mylącą antropomorfizację komputerów. Trzeba jednak przyznać, że podobny zarzut można również postawić wobec innych zbliżonych określeń, takich jak „konfabulacja" czy „fabrykowanie".

Niezależnie od przyjętego określenia, warto powtórzyć, że twórcy platform SI zdają sobie sprawę ze skali wyzwania i stosują obecnie „bezpieczniki" ograniczające halucynacje. Nowsze modele LLM zostały nauczone, aby udzielać odpowiedzi „nie wiem" na każdy prompt. Jak w takim razie cokolwiek generują? Uruchamia się wtedy rodzaj generatywnej sieci GAN (*Generative Adversarial Network*), której dwie części konkurują ze sobą – jedna wytwarza odpowiedzi (generator), druga ocenia, na ile wytworzone dane przypominają rzeczywiste (dyskryminator). Jeżeli dyskryminator uzna, że odpowiedź przypomina prawdę, wyłącza moduł zabezpieczający.

Myśli czarnej skrzynki

Skoro tak, to dlaczego SI halucynuje? Internet jako źródło wiedzy jest generalnie mało wiarygodny – wypełniony jest przecież nie tylko rzetelnymi materiałami źródłowymi, lecz również opiniami, przeinaczeniami i dezinformacją. Modele sztucznej inteligencji nie potrafią same ocenić, czy konkretne sformułowania są prawdziwe (mogą za to ocenić, czy przypominają prawdę) – zwłaszcza wówczas, gdy są pytane o szczegóły dotyczące słabo opisanego tematu.

Chatboty łączą najrozmaitsze fragmenty tekstów z tysięcy, jeśli nie milionów źródeł, „wypluwając" nowe materiały. Co więcej – zwykle na to samo pytanie zadane dwa razy uzyskuje się dwie różne odpowiedzi (czasem uzupełniające się, a czasem zaprzeczające sobie). Eksperci starający się wyjaśnić problem halucynacji sztucznej inteligencji, jak na przykład ekonomista i statystyk Gary Smith (który zresztą nie kryje swojego krytycznego stosunku do SI), podkreślają, że algorytmy nie rozumieją słów, a zatem nie mogą ocenić, czy analizowane i generowane ciągi znaków opisują rzeczywistość, czy są fałszem.

Inna sprawa, że platformy SI, których celem istnienia jest odpowiadanie na pytania użytkowników – po prostu to robią. W przypadku dużych modeli językowych (LLM) typu GPT (*Generative Pre-trained Transformer*) sztuczna inteligencja trenowana jest, by przewidywać następne słowo w szyku. I to słowo się pojawia – nawet wówczas, gdy SI nie ma wystarczających danych pozwalających wygenerować prawdziwą odpowiedź. W przypadku innych typów SI problemem może być źle dobrany zestaw danych treningowych, które utrwalają błędne „skojarzenia". Jednak rzeczywistym wyzwaniem leżącym u podstaw halucynacji SI jest brak wyjaśnialności sztucznej inteligencji, czyli zrozumienia, dlaczego model podjął określoną decyzję.

Jak pokazują dane opublikowane w październiku ubiegłego roku na łamach magazynu „Nature", chatboty działające na najpopularniejszych platformach SI mają tendencję do udzielania takich odpowiedzi, które satysfakcjonują użytkownika – a niekoniecznie są prawdziwe. Zespół Myry Cheng z Uniwersytetu Stanforda sprawdził zachowania 11 chatbotów, w tym ChatGPT, Gemini, Claude, Llama i DeepSeek. Okazało się, że krzemowi pochlebcy dopasowywali swoje odpowiedzi do oczekiwań ludzi znacznie częściej, niż zrobiliby to żywi rozmówcy.

I tu pojawia się kolejny problem z halucynacjami SI – często są one wynikiem sprytnie napisanego promptu (przykład niedźwiedzi w kosmosie). Może się on odnosić do nieistniejącej osoby, zdarzenia, które nigdy nie miało miejsca, albo „naciskania" na model językowy, aby udzielił satysfakcjonującej odpowiedzi.

Przywidzenia kierowcy

Halucynacje modeli SI – podobnie jak u ludzi – mogą być zabawne, ale mogą też przynosić szkody. Tak było m.in. z wpisami Groka (systemu SI w serwisie X) negującymi Holokaust i wybielającymi Niemców. Sztuczna inteligencja napisała m.in., że krematoria w niemieckim obozie Auschwitz zostały pierwotnie zaprojektowane jako instalacje dezynfekcyjne do zwalczania chorób zakaźnych. Trzeba tu podkreślić, że Grok jest bardzo często wykorzystywany do weryfikowania prawdziwości twierdzeń użytkowników X (dawniej Twitter), a jego odpowiedzi praktycznie kończą dyskusję.

Kategoria „niebezpieczne konfabulacje" to niestety nie tylko chatboty. Trzeba tu wspomnieć np. o systemach SI analizujących obraz i kierujących pojazdami autonomicznymi. System, który zobaczy na ulicy nieistniejącego psa, może spróbować gwałtownie skręcić, aby uniknąć wypadku. Jeden z pierwszych, a na pewno niestety najsłynniejszy wypadek samochodu Tesla jadącego na Autopilocie w maju 2016 roku zdarzył się, ponieważ algorytm niewłaściwie ocenił otoczenie – „nie zauważył" białej ciężarówki na tle jasnego nieba – halucynował, że droga jest pusta.

To samo dotyczy niewłaściwej interpretacji wyników badania obrazowego. Obecnie, również w Polsce, oszałamiającą karierę robi teleradiologia – specjalista oceniający np. wynik badania tomograficznego czy rezonansu magnetycznego otrzymuje przez internet tylko obrazy, które musi zinterpretować. Często taki specjalista wspomaga się wyspecjalizowanym oprogramowaniem bazującym na sztucznej inteligencji. Ta zaś, na co zwracają uwagę sami

lekarze zlecający badania, potrafi pomylić skrzep krwi z ogniskiem nowotworowym. Podobny błąd może popełnić SI analizująca zmiany skórne – zaalarmować pacjenta z powodu nieistniejącego czerniaka.

Są też takie obszary, w których ludzie zapewne nigdy nie dostrzegą halucynacji. Doskonałym przykładem może być rynek finansów i ubezpieczeń. Trudno tu zweryfikować decyzje sztucznej inteligencji obliczającej wysokość składki, kalkulującej ryzyko czy przygotowującej plany biznesowe.

Jeszcze jesteśmy potrzebni

Jak często halucynuje sztuczne inteligencja? Od dwóch lat sprawę zmyślających chatbotów bada firma Vectara. Okazuje się, że nawet przy prostym zadaniu podsumowania artykułów chatboty konfabulowały – w zależności od modelu wymyślały od 3 proc. do nawet 27 proc. treści. Opisywane wyżej obwody zabezpieczające sprawiają, że nowe platformy LLM oszukują rzadziej – w przypadku modeli Google i OpenAI halucynacje to zaledwie 1–2 proc. odpowiedzi, a Anthropic – ok. 4 proc.

Co ciekawe, wskaźniki te pogarszają się u najnowszych tzw. modeli rozumujących (RLM). DeepSeek R1 ulega halucynacjom w ponad 14 proc. odpowiedzi, a OpenAI o3 – w prawie 7 proc. W niektórych testach te najbardziej zaawansowane modele sztucznej inteligencji podawały nieprawdziwe odpowiedzi na blisko co drugie pytanie. Dlaczego? Modele rozumujące są zaprojektowane tak, aby mogły poświęcić więcej czasu na „przemyślenie" złożonych problemów przed znalezieniem odpowiedzi. Rozwiązują problem krok po kroku, co sprawia, że narażają się na ryzyko halucynacji na każdym etapie. Im więcej czasu poświęcają na „rozumowanie", tym większe ryzyko halucynacji.

Wygląda zatem na to, że problemu halucynacji sztucznej inteligencji nie da się rozwiązać w prosty sposób – wynika on bowiem z samej jej zasady działania. W tym wszystkim jest jednak dla nas, ludzi, dobra wiadomość: nawet w erze SI liczy się rzetelna wiedza, zdrowy rozsądek i sporo sceptycyzmu wobec nowych technologii. Jak to ujmuje prof. Subbarao Kambhampati, badacz relacji SI – człowiek na Uniwersytecie Stanowym Arizony: „jeśli jeszcze nie znasz odpowiedzi na pytanie, to raczej nie zadawaj tego pytania sztucznej inteligencji".

Jak sobie radzić z oszukującą sztuczną inteligencją?

O to, jak uniknąć błędów, zapytaliśmy jednego z najlepszych ekspertów w tym temacie. Któż bowiem może wiedzieć więcej o halucynacjach SI niż sama SI, a konkretnie ChatGPT.

Halucynacje AI (czyli wymyślanie faktów, źródeł lub odpowiedzi „brzmiących pewnie, ale nieprawdziwych") da się znacznie ograniczyć, choć nie da się ich wyeliminować w 100 proc. Oto praktyczne i sprawdzone sposoby:

1. **Zadawaj precyzyjne pytania**
   Im bardziej ogólne pytanie, tym większe ryzyko halucynacji. Proś o daty, numery dokumentów, autorów.

2. **Wymagaj źródeł (i je sprawdzaj)**
   Zawsze proś o linki do oficjalnych stron, publikacji naukowych, aktów prawnych. Jeśli SI nie potrafi podać źródła albo podaje je ogólnikowo („badania pokazują…"), to sygnał ostrzegawczy.

3. **Ogranicz zakres odpowiedzi**
   Poproś, aby SI odpowiadała tylko na podstawie znanych danych i jasno zaznaczała niepewność.

4. **Unikaj pytań „zmyśleniowych"**
   SI najczęściej halucynuje przy bardzo niszowych osobach lub wydarzeniach, nieistniejących publikacjach, pytaniach typu „czy pamiętasz dokument X z 2014 r.?". Warto zapytać „czy istnieją wiarygodne informacje na ten temat?"

5. **Sprawdzaj kluczowe fakty w drugim źródle.**
   Traktuj SI jak asystenta, nie eksperta. Jest to szczególnie ważne w medycynie, prawie, finansach, historii i datach.

6. **Używaj SI do struktury, nie do faktów**
   Sztuczna inteligencja świetnie streszcza, porządkuje informacje, generuje checklisty i tłumaczy tekst. Ale fakty najlepiej weryfikować samodzielnie.

7. **Zadawaj pytania kontrolne**
   Poproś SI, aby wypunktowała założenia, wskazała, co może być niepewne i podała alternatywne interpretacje.

8. **Ustaw „tryb ostrożny" w promptach**

Odpowiedź ChatGPT została edytowana w taki sposób, aby była czytelna i spójna z formą pisma „Domena" – red.